OPEN ACCESS

University of BRISTOL

Gubian, M., Davis, C., Adelman, J., & Bowers, J. (2017). Comparing single-unit recordings taken from a localist model to single-cell recording data: a good match. *Language, Cognition and Neuroscience*, *32*(3), 1-33. https://doi.org/10.1080/23273798.2016.1259482

Peer reviewed version

Link to published version (if available):
10.1080/23273798.2016.1259482

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Taylor & Francis at http://www.tandfonline.com/doi/full/10.1080/23273798.2016.1259482. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Comparing single-unit recordings taken from a localist model to single-cell recording data: A good match.

Michele Gubian[1]

Colin J. Davis[1]

James S. Adelman[2]

Jeffrey S. Bowers[1]

1. School of Experimental Psychology, University of Bristol, UK

2. Department of Psychology, University of Warwick, UK

Corresponding author:

M.Gubian , e-mail: mm14722@bristol.ac.uk

Abstract

Single-cell recording studies show that some neurons respond to complex visual information (e.g., words, objects, faces) in a highly selective manner, with individual neurons responding to about .5% of presented images. Such data have often been taken as inconsistent with "grandmother cell" theories as well as with localist models in psychology. In particular, it is commonly assumed that units in localist models respond to only one input, resulting in greater levels of selectivity than seen in single-cell results. To test this assumption, we recorded unit activity from a localist model of word identification. Our results show that the model can capture the levels of selectivity reported in neuroscience. Accordingly, single-cell data do not rule out localist coding schemes. We propose that the term grandmother cell should be reserved for the hypothesis that the brain implements localist representations: neurons that represent one and only one thing but respond to multiple things.

Keywords: localist coding, grandmother cells, selectivity, sparseness, computational models, visual word recognition

Introduction

There are many ways that "grandmother cells" can be defined so that they are easily falsified. For instance, if a theory of grandmother cells entails the view that each of our visual experiences (indeed all possible future experiences) is coded by single neurons, then it is clear this view is untenable because there are not enough neurons in a brain to code for all possible different experiences (e.g., Eichenbaum, 2001; Plaut & McClelland, 2010). Similarly, if a grandmother cell theory entails the view that one and only one neuron codes a given object (i.e., there is no redundancy in coding), then it is subject to the criticism that losing a single neuron could lead to the selective loss of knowledge (e.g., Lytton, 2007).

In the same way, if a grandmother cell theory entails the view that neurons selectively respond to one thing *and nothing else* (that is, baseline firing to all other stimuli), then it is hard to explain the frequency with which researchers have identified highly selective neurons (e.g., Waydo, Kraskov, Quiroga, Fried, & Koch, 2016; Yuste, 2015). That is, the probability of finding a grandmother cell of this sort should be extraordinarily small given that researchers can only record from a few dozen or at most a few hundred neurons (out of millions), and given that researchers can only present a tiny fraction of the possible images to a participant. So, counter-intuitively, the frequent reports of highly selective neurons are taken as evidence against grandmother cells. Equally problematic for this view, even the most selective neurons tend to respond to other images (e.g., a neuron in hippocampus that selectively responded to images of Jennifer Aniston also fired to an image of Lisa Kudrow, her co-star on the television show *Friends;* Quiroga, Reddy, Kreiman, Koch, & Fried,

2005). The common conclusion from the above observations and results is that grandmother cell theories are implausible. This in turn is used to argue that knowledge is coded in a distributed format, with many neurons involved in coding a given stimulus, and each neuron coding many different inputs.

The problem with this line of argument is that these observations are also consistent with localist coding, the main theoretical alternative to distributed coding within cognitive psychology. In simulations below we show that a localist model of visual word identification can account for the frequency with which selective units are found and the observation that many highly selective neurons often fire to more than one input (in this case, words). These findings are important because many theorists reject localist models in psychology on the basis of the neuroscience.

Localist vs. distributed coding in cognitive psychology

Within psychology localist and distributed models have been advanced across a wide range of domains, including visual and spoken word identification, short-term memory, episodic memory, semantic memory, object perception, face perception, motor control, etc. On localist theories, individual words, objects, simple concepts, etc., are coded distinctly, with their own dedicated representation. For example, the words FOG and DOG would be coded with distinct and non-overlapping mental representations. On any localist account, the words FOG and DOG would be linked by virtue of sharing some features (e.g., letters), but the words themselves would be stored explicitly and separately in the mind. For example, the Interactive Activation (IA) model of visual word identification (McClelland & Rumelhart, 1981) includes localist representations for letter features, letters, and words, and word identification is achieved when a single unit is activated beyond some threshold. As a consequence, it is possible to record from a single unit and determine whether the model is

processing a given word or not.

The use of localist representations was challenged by the introduction of Parallel Distributed Processing (PDP) models (McClelland, Rumelhart, & PDP Research Group, 1986; Rumelhart, McClelland, & the PDP Research Group, 1986). A key claim of this approach is that knowledge is coded in a distributed manner in the mind and the brain. That is, knowledge is coded as a pattern of activation across many processing units, with each unit contributing to many different representations. A classic example is the developmental model of word recognition and naming by Seidenberg and McClelland (1989) that can name many familiar (both regular and irregular) and unfamiliar (untrained) words on the basis of learned distributed codes. As a consequence, even though the model can succeed in naming the word DOG, it is not possible to determine the identity of the word by recording the activation of a single unit in the hidden layer.

Critically, proponents of PDP models often cite evidence from neuroscience in support of distributed models compared to localist models. For example, McClelland and Ralph (2015) write:

> Neuronal Recording Studies relying on microelectrodes to record from neurons in the brains of behaving animals can allow researchers to study the representations that the brain uses to encode information and the evolution of these representations over time. Several fundamental observations have been made using this technique. As discussed above, these studies indicate, among other things, that the brain relies on distributed representations…

Similarly, Flusberg and McClelland (2014) wrote:

> The key take-home message is that connectionist models are not just somewhat biologically plausible implementations of existing psychological theories; rather, they

> provide alternatives to other theories and offer a means of investigating a unique way of thinking about mental processing….

This last quote not only highlights the common claim that PDP models are more biologically plausible than alternative (localist) approaches, but in addition, highlights the claim that this is an important theoretical distinction.

In fact we agree that localist and distributed coding schemes constitute important alternative theories of how knowledge is represented (cf. Bowers, Vankov, Damian, & Davis, 2014, 2016; Page, 2000). The question we address here is whether indeed single-cell recording data do indeed rule out localist coding schemes. We assess this by carrying out single-unit recordings in a localist model of visual word identification (Davis, 2010) and comparing the results with single cell studies that are typically taken as inconsistent with grandmother cell coding, as we detail next.

Comparing the selectivity of neurons and localist units in a model of word identification

Some of the most striking demonstrations of selective neuronal responding have been reported in the medial temporal lobes of human participants. For example, Quan Quiroga et al. (2005) recorded from a total of 993 units (single neurons or small groups of neurons) from eight patients with epilepsy where electrodes are implanted in order to localize the focus of the seizure. The patients were tested over the course of 21 sessions with approximately 90 images (photographs of people, objects, and scenes) presented per session. They identified 132 units (14%) that responded to at least one picture, and all these responses were highly selective, with responsive neurons only responding to 2.8% of the presented pictures (range: .9-18%). Furthermore, 51 of the responsive units (38.6%) respond to different photos of the same person, building, animal, or object (they showed some invariance). For instance, one

neuron responded to multiple images of Jennifer Aniston (star of the TV show *Friends*).

Although these findings might be taken as lending some support to grandmother cells, the authors ruled out this interpretation based on two considerations. First, although some neurons only responded to one category of image within the experimental session (for instance, one neuron responded to multiple images of the actor Steve Carell and was near silent to 54 other faces), other neurons responded to different categories of images (for instance the Jennifer Aniston neuron also responded to an image of Lisa Kudrow -- a co-star of the TV series *Friends* -- and another neuron responded to Tower of Pisa and the Eiffel Tower). The authors concluded that all neurons would be activated by multiple different categories of images if more images were presented in a given session (technical constraints restricted a given session to about 100 images). Second, the fact that these neurons were found in the first place (out of the many millions of neurons in MT) implies that the neurons must respond to a wide variety of different images (otherwise they would not be found). A subsequent analysis of these results (Waydo, Kraskov, Quiroga, Fried, & Koch, 2006) estimated that the average selectivity of these neurons was approximately .5%, meaning that these neurons responds to about .5% of presented images, and that each neuron responds to between 50-150 different categories of images (on the assumption that adults recognize between 10,000 and 30,000 discrete objects)[Footnote1]. This then provides an explanation of the frequency with which they reported these highly selective cells in previous experimental sessions. They noted that if one assumes .54% selectivity, recording from an average of 54 neurons per session and presenting 88 images should result on average in finding 15.9 units responding to 17.9 stimuli. This is not so far from what they in fact observed in each session.

Again, we agree with the authors that these findings are hard to accommodate with the view

that a single grandmother neuron codes for a given person, place, or thing, and that grandmother neurons only respond to one thing (being entirely silent to everything else). But for present purposes, the question is whether these findings are problematic for localist coding models in psychology.

To assess whether indeed these findings are problematic for localist theory we analysed the word units in the Spatial Coding Model (SCM) of visual word identification (Davis, 2010) that represents words with localist codes. We carried out simulated single-unit recording studies in which we recorded the activity of word units while presenting a random sample of words to the model. The question is whether our single unit recording results are broadly consistent with the single-neuron recording results reported in the brain that are taken to be inconsistent with localist coding.

Simulation 1

The SCM is a localist model of visual word recognition that can simulate a broad range of findings in visual word recognition (e.g., Davis, 2010; Lupker, Zhang, Perry, & Davis, 2015; Stinchcombe, Lupker, & Davis, 2012). Published simulations of the model have used a vocabulary of 30,605 English words, comprising all words of between two and ten letters with frequencies of occurrence of greater than 0.34 per million in the Celex English Corpus Types corpus (Baayen, Piepenbrock, & van Rijn, 1995). This number of words is in the same range as the number of discrete objects that a typical adult can recognise, according to the estimate by Biederman (1987), as cited by Waydo et al. (2006). The size of this model's vocabulary was our principal reason for choosing this particular localist model, rather than other localist models such as the IA model (which is restricted to 1,179 4-letter words) or the DRC model (which is restricted to 7,954 monosyllabic words).

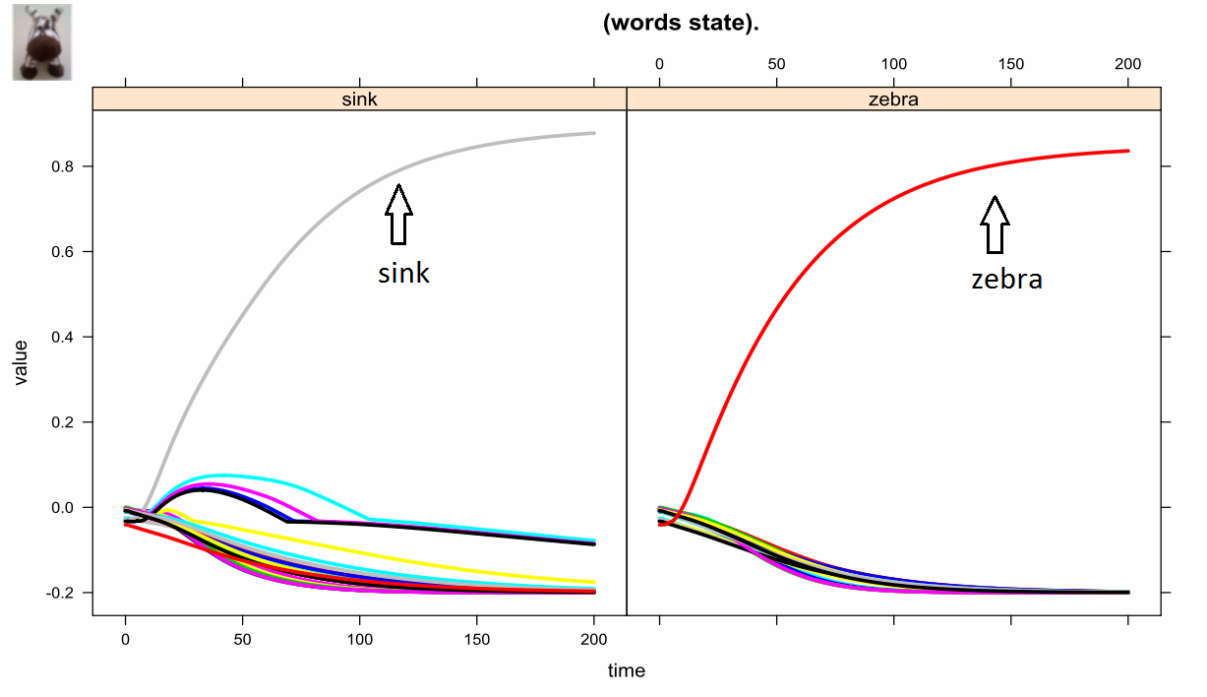

Figure 1: Example activation functions over time in the original SCM for two different word stimuli (*sink* and *zebra*).

When a familiar word is presented to the SCM (i.e., a word contained in the model's vocabulary) activity in the corresponding word unit shows a sigmoid activation function over time *y(t)*, in which there is a brief interval where activity *y* increases faster than linearly ($\frac{\partial y(t)}{\partial t} > 0, \frac{\partial^2 y(t)}{\partial t^2} > 0$), followed by an interval where activity grows roughly linearly ($\frac{\partial y(t)}{\partial t} > 0, \frac{\partial^2 y(t)}{\partial t^2} \cong 0$), before finally converging towards an equilibrium state at a slower than linear rate ($\frac{\partial y(t)}{\partial t} > 0, \frac{\partial^2 y(t)}{\partial t^2} < 0$). Typical activation functions are shown in Figure 1 for the cases where the word stimulus is either *sink* or *zebra*. In the latter example, there are no words in the vocabulary that are close orthographic neighbours of the word *zebra*, and hence the *zebra* word unit is the only word unit that shows any increase in activity over time; for the remaining word units, activity declines towards the lower limit (of course, there are other units in the model that do become active, such as letter units and letter feature units, but we

focus here on the word layer, which constitutes the vast majority of the units in this model). The situation is slightly different for the word *sink*: although the activation function for the winning word unit is very similar to that seen in the previous example, here we can see some evidence of other word units beginning to activate before lateral inhibition from the *sink* word unit has the effect of suppressing their activity. Those units which become partially activated are orthographic neighbours of *sink* — *sick, silk, sank* and *sunk*. This example already illustrates that activity is not limited to a single grandmother cell representation, i.e. activity is not as selective as in the straw man (or straw grandmother) characterizations of localist networks. Nevertheless, it is the case that, for this particular model, relatively few units are driven above their baseline activity by familiar word stimuli.

Method

We performed a series of quantitative analyses aimed at assessing the compatibility of results from single-neuron recording studies with computational models based on localist networks. To make the case concrete, we attempted to compare the behaviour of the SCM (Davis, 2010) with the probabilistic analysis of single-neuron recording studies reported by Waydo et al. 2006. Units in the SCM take on activation values that abstracts away from the complexity of physical neurons. As a consequence, there is no obvious one-to-one mapping between physical measurements in terms of neuron spikes per second with model unit activation in time. To make the comparison between Waydo et al. 2006 and SCM possible we made a number of choices and assumptions (see Table 1). Given these assumptions, the comparison becomes relatively straightforward.

Waydo et al. 2006	**SCM**
"we considered a response to be significant if it was larger than the mean plus a threshold number of SDs of the baseline (before the onset of the image) and had at least two spikes in the poststimulus time interval considered"	There is no analog to spikes. We consider a response to be significant, i.e. a unit *responds*, if its level of activation rises at least $\theta_y \geq 0$ above its starting (resting) level during stimulus presentation. (See text on the choice of θ_y)
"poststimulus time interval considered (0.3–1 s)"	Time interval is taken to be a number of simulation cycles T after stimulus presentation largely sufficient for the model to perform lexical identification.
Selectivity is estimated by looking at the (estimated) distribution of its value given responses of multiple neurons to stimuli presentation.	Selectivity is estimated by counting the number of responses N_R each unit provides across the presentation of all $N = 30{,}605$ word stimuli known to the model. The estimator for selectivity is the median of N_R/N across all the N units in the word layer. (Note that the number of units N in the word layer coincides with the number of words known to the model)
Stimuli are images	Stimuli are written words

Universe of stimuli cardinality assumed to be around 30,000	Universe of stimuli cardinality known exactly and equal to the model vocabulary size, i.e. the number N of word units, $N = 30{,}605$.
Number of neurons unknown	Number of units known and equal to the universe of stimuli cardinality, $N = 30{,}605$.
"A large majority of neurons within the listening radius of an extracellular electrode are entirely silent during a recording session. … Thus, the true sparseness could be considerably lower."	All unit activities are recorded.
"there is a sampling bias in that we present stimuli familiar to the patient (e.g., celebrities, landmarks, and family members) that may evoke more responses than less familiar stimuli."	We avoid this potential bias by selecting words at random from the set of words known to the model.

Table 1: Correspondence between criteria used to identify selective neurons in single-unit recording studies and criteria we used in our simulations.

All of the simulations reported in this article were conducted using the *easyNet* simulation

software developed by Adelman, Gubian and Davis (2016)[Footnote 2][Footnote 3]. The estimate of selectivity in SCM was computed as follows. All $N = 30{,}605$ words in the model vocabulary were independently presented to the model for a fixed amount of time $T = 100$ simulation cycles. The model did not perform any task; we simply recorded word unit activities. A unit is said to *respond* to a stimulus if at any point in time t during stimulus presentation ($0 \leq t \leq T$) the following holds true:

$$y(t) - y(0) > \theta_y \geq 0 \quad (1)$$

where $y(t)$, a value between -0.2 and 1.0, is the level of activation at each point in time (here t is a discrete value that indexes simulation cycles), and θ_y is a threshold to be specified. In this way each unit provides a binary outcome at the end of a stimulus presentation, i.e. it did or did not respond to the stimulus. This simple definition was chosen after observing the typical system dynamics, whereby at the start of a stimulus presentation a number of units raise to some level above their resting or baseline level, i.e. the value $y(0)$, and subsequently decrease again due to competition (except for the winner), while the large majority of units steadily decrease right from the start (cf. Figure 1). Selectivity is estimated from the observations of the quantity N_R/N, where N_R is the number of responses (as defined in Eq. (1)) a single unit gives following the presentation of N stimuli. Since the distribution of N_R/N across the $N = 30{,}605$ units is usually quite skewed, we will adopt the median as estimator.

Results and Discussion

Figure 2 reports empirical cumulative distribution functions (ECDFs) of the number of responses N_R as defined above. The different functions represent different choices for the threshold θ_y (cf. Eq.(1)). For example, the curve corresponding to $\theta_y = 0$ shows the number

of stimuli for which a unit responded by increasing its activity above its resting level. By reading the median off the curve (i.e. the value of N_R corresponding to ECFD = 0.5) we note that this value is 19, meaning that typically units would go above their baseline activity $y(0)$ only for 0.06% of stimuli ($\frac{19}{30506}$), or in other words, units do not go above their baseline activity in 99.94% of the cases. This shows that it is safe to consider as a response even an increase in activity above baseline level greater than say $\theta_y = 0.01$, or to be more conservative, $\theta_y = 0.05$. Figure 2 shows that for those values of θ_y medians of N_R are 4 and 2 for $\theta_y = 0.01$ and $\theta_y = 0.05$, respectively. Even adopting the more inclusive criterion $\theta_y = 0.01$ we get an estimated selectivity of 0.013%, which is far smaller than the 0.5% selectivity rate estimated by Waydo et al. (2006).

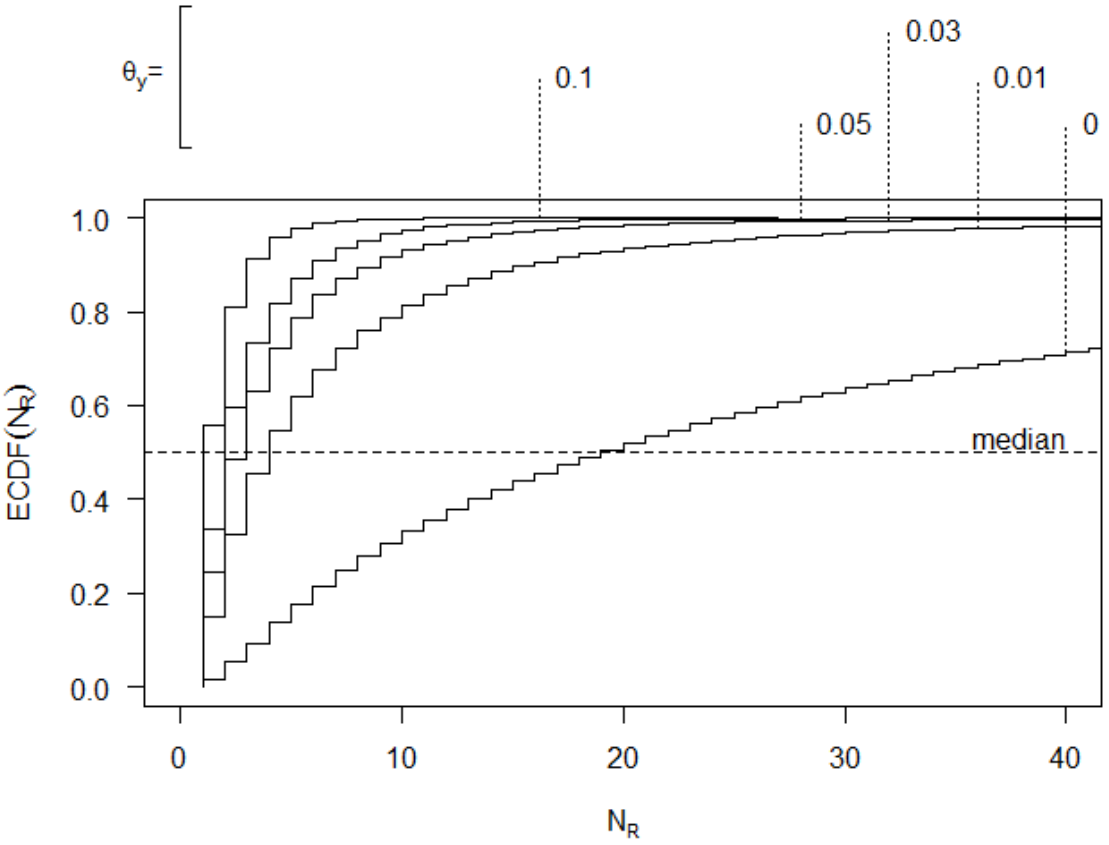

Figure 2: Empirical cumulative distribution functions (ECDFs) of number of responses N_R for several values of activation threshold θ_y (cf. Eq. (1)) obtained by presenting $N = 30{,}506$ words to SCM for $T = 100$ cycles. All parameters of SCM are as in the published model (Davis, 2010).

Simulation 2

The extreme selectivity observed for the SCM model in Simulation 1 is a problem for the model not simply from the perspective of matching data from single cell recording studies, but also for explaining behavioral results. For example, using a semantic categorisation task, Bell, Forster, and Drake (2015) found evidence for priming relationships between relatively orthographically distant words (e.g., *capable* priming *cabbage*, *senior* priming *sailor*). Bell et al. refer to such pairs as "SCM neighbours" on the basis that they were chosen on the basis of

their moderate similarity, as measured by the match scores computed in the SCM model. The present version of the model cannot capture such priming effects because such orthographic neighbours are not typically among the relatively small set of word units that become activated by the prime. That is, the lexical activation in the model is somewhat too selective to support such priming.

It is important to note, though, that selectivity is not a defining feature of localist neural network models. Mathematical analyses of the relevant properties of these networks have been described by Grossberg (1973), who noted the existence of a *quenching threshold*, below which input signals are treated as noise and suppressed. Inputs greater than the quenching threshold are preserved or amplified, depending upon model parameters and the nature of the lateral inhibition (including the signal function). Thus, modifying the lateral inhibitory function of the original spatial coding model can result in changes to the selectivity of activity at its word layer. In particular, Davis (1999, chapter 2) noted that the approximately winner-take-all behaviour associated with models like interactive activation (McClelland & Rumelhart, 1981) can be transformed to a less selective coding at the word layer if the lateral inhibitory connection is replaced with what Grossberg refers to as *shunting inhibition*, whereby the inhibitory input to each unit is weighted by that unit's current activity (specifically, the inhibitory signal becomes magnified as a unit's activity increases). This enables word units which provide a moderate match to the input stimulus to begin to become active without being immediately suppressed by the best matching word unit. This form of inhibition is already used in other layers in the spatial coding model, but was eschewed at the word layer in the published model so as to permit a generalisation of the original interactive activation model (which does not use shunting inhibition). However, it is possible to rewrite the model's equation for computing lateral inhibition to allow a generalisation that can

encompass both shunting and non-shunting inhibition models. That is, we can replace the original expression for the strength of lateral inhibitory signals:

$$Q_i = \lambda_1(\chi - [x_i]^+) \quad (2)$$

with a generalised form which allows for self-shunting inhibition:

$$Q_i = (\lambda_1 + \lambda_2\ [x_i]^+)\ (\chi - [x_i]^+) \quad (3)$$

Here, Q_i represents the lateral inhibitory input to the ith word unit, x_i represents the current activity of the ith word unit, χ represents the summed activity of all word units (weighted according to the length of the words they code; see Davis (2010), eq. 30), and the notation $[x_i]^+$ indicates a function that applies a lower bound of zero to unit activities (e.g., units with negative activities do not result in a negative value of Q_i). In equation (3), setting λ_1 to .34 and λ_2 to 0 is equivalent to the original SCM. On the other hand, setting parameter λ_1 to 0 and λ_2 to some positive value results in a model in which lateral inhibition at the word level depends entirely on shunting inhibitory signals. This is the version of the model that we test in Simulation 2, adopting a setting of $\lambda_2 = .3$. With this modification to inhibition in place, the selectivity of the model depends largely on the bottom-up letter word excitation strength [Footnote 4], which effectively determines the quenching threshold. We also modified two other parameters: a) the y_{global} parameter (which weights the contribution of total lexical activity to "Yes" responses in the lexical decision task; see equation 39 in Davis, 2010) was reduced from .4 to .04; this parameter adjustment was necessary to compensate for the greater level of total lexical activity in the modified model; b) the γ_{len} parameter (which weights the strength of inhibition to word units that differ in length from the number of letters in the current stimulus; see equation 34 in Davis, 2010) was increased from .06 to .3; this parameter change reflects the decreased reliance on lateral inhibition to "clean up" the input.

Method

The methodology for presenting stimuli to the network and measuring selectivity was the same as in Simulation 1.

Results and Discussion

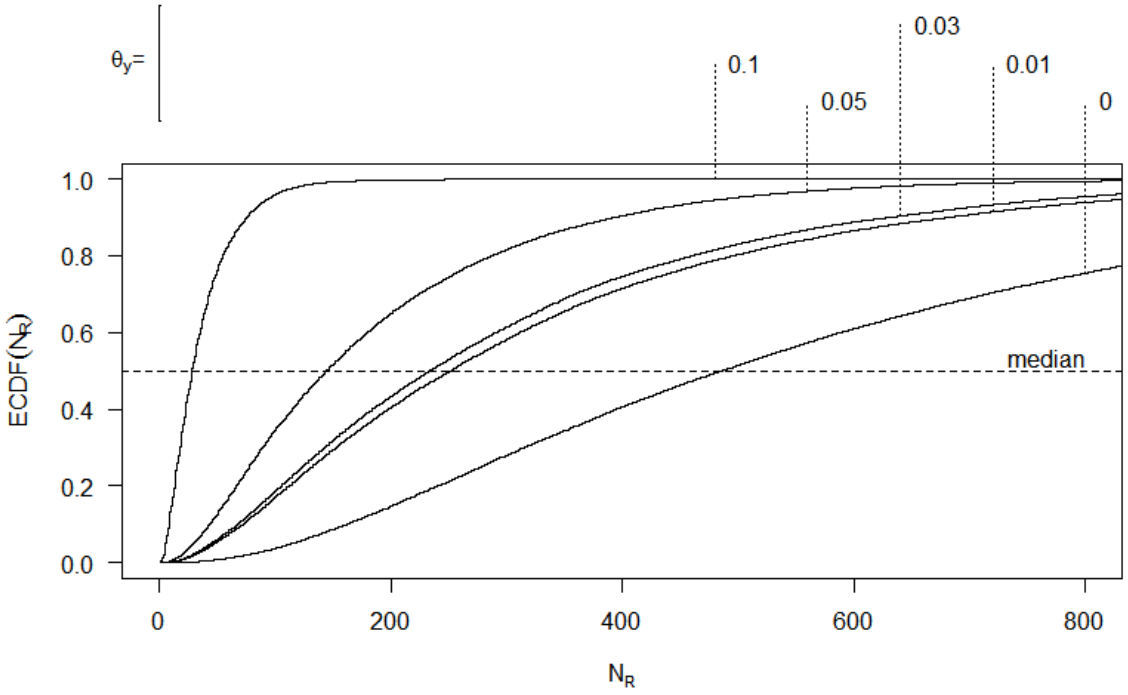

Figure 3: Empirical cumulative distribution functions (ECDFs) of number of responses N_R for several values of activation threshold θ_y (cf. Eq. (1)) obtained by presenting $N = 30{,}506$ words to the modified SCM for $T = 100$ cycles.

Figure 3 shows the empirical cumulative distribution functions of network selectivity given a range of θ_y values with a value for bottom-up letter word excitation strength $\alpha = 2.0$. Comparison with Figure 2 indicates that the network using shunting inhibition produces activity that is still highly selective. The curve for $\theta_y = 0$ has its median value at $N_R = 486$, which means that typically a unit responds to only 1.6% (i.e., $\frac{486}{30506}$) of the word stimuli

belonging to the model's vocabulary. Nevertheless, this outcome is considerably less selective than the original model. A conservative choice for the threshold $\theta_y = 0.05$ provides an estimate of selectivity of 0.47% (median $N_R = 144$), which closely matches the 0.5% estimate by Waydo et al.

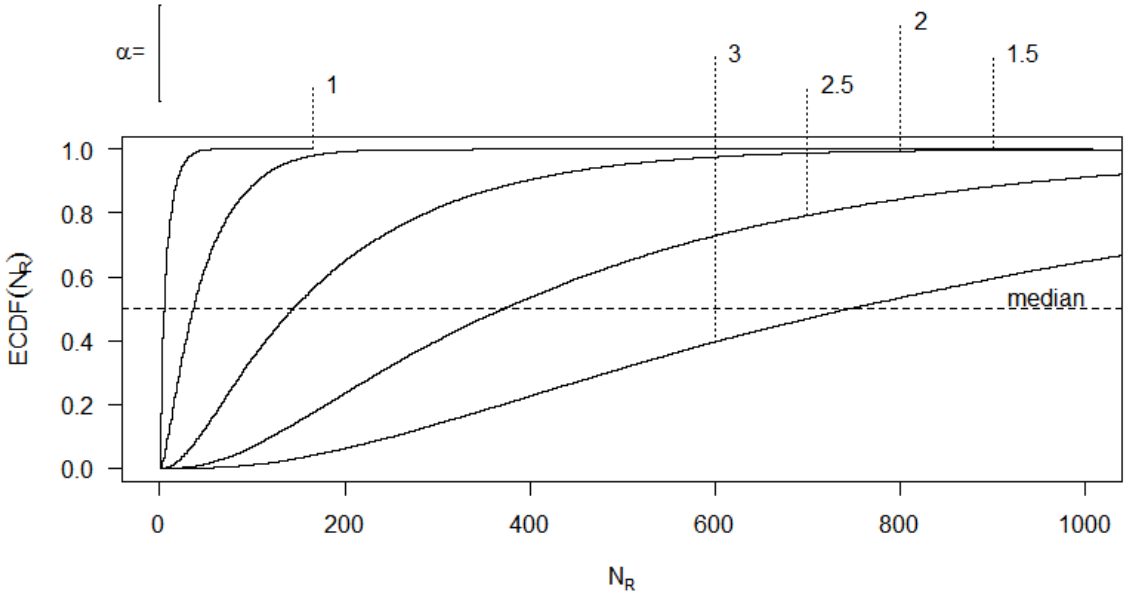

Figure 4: Empirical cumulative distribution functions (ECDFs) of number of responses N_R for several values of α (parameter scaling excitatory input to the word layer) and for a fixed value of θ_y =0.05.

Figure 4 shows how the empirical cumulative distribution functions vary as a function of the value of the excitatory input parameter α while keeping $\theta_y = 0.05$. As can be seen, there is a clear relationship between the strength of the excitatory input and the selectivity of the network – in principle, it is possible to achieve any desired level of selectivity by varying the magnitude of α. Thus, there is no sense in which results from single cell recording studies demonstrate levels of selectivity that are inconsistent with localist models. This observation does not render localist models unfalsifiable, but it does imply that falsification cannot be

achieved purely on the basis of data from single cell recording studies (though such data could constrain the parameterisation of models, and in combination with behavioural results could falsify specific models).

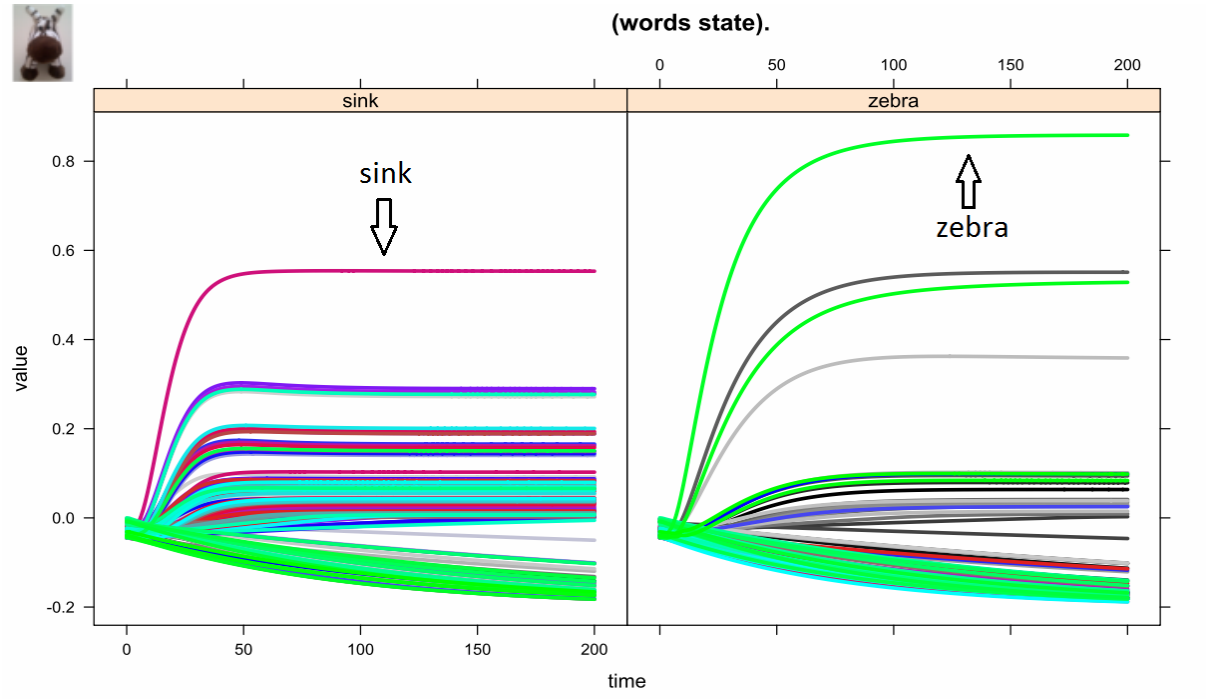

Figure 5: Example activation functions over time in the modified SCM for two different word stimuli (*sink* and *zebra*).

The effect of the changes to the model can also be seen by examining activity over time. Figure 5 shows the activity functions for the same two word stimuli as in Figure 1. Comparing the two figures reveals a couple of important differences. First, there are clearly many more word units activated in the modified model than in the original model (nevertheless, it remains the case that over 98% of word units show no positive response to these stimuli). The second difference is that the equilibrium activity of the winning unit is considerably lower when the stimulus is *sink*, which has many orthographic neighbours, than when the stimulus is *zebra* (which activates far fewer units). This difference reflects the

effect of lateral inhibition, which effectively conserves the total activity at the word layer.

Of course, it is important to establish that the modified model can still provide a good model of visual word recognition. A comprehensive analysis of the behaviour of this model is beyond the scope of the present article; simulating the same set of phenomena as the original SCM model would require careful exploration of the parameter space of the model. Here we simply note that the model satisfies the fundamental property required of a model of visual word identification, i.e., it can identify familiar words. To test this, we sampled 1000 words and 1000 nonwords from the set of stimuli contained in the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012). The sampling criteria were that stimuli were classified correctly by >99% of participants and did not contain common suffixes (-ing, -er, -ed, -es and -s); the latter criterion reflects the theoretical assumption, motivated by experimental findings (e.g., Marslen-Wilson, Tyler, Waksler, & Older, 1994), that recognition of morphologically transparent words involves a decomposition process (e.g., *swelling* → *swell* + *-ing*) that is not specified in the current model. Each of the 1000 words was correctly identified by the model (the identification threshold was set at 0.35). When the task was lexical decision (i.e., categorising stimuli as words or nonwords), the model correctly classified 100% of the words and 98% of the nonwords (threshold activities for the YES and NO channels were set to .35 and .3, respectively). The nonwords that were miscategorised as words were very wordlike nonwords such as *twetty* or pseudocompounds such as *whiskwince*. Thus, despite the much greater level of lexical activity, the model is near-perfect at identifying words and discriminating words from nonwords.

General Discussion

It is widely claimed that single-cell recording data in neuroscience support distributed

as opposed to localist models in psychology (McClelland & Ralph, 2015; O'Reilly, 1998; Plaut & McClelland, 2010). A key problem with this claim, however, is that it has not been tested. Indeed, in contrast with the vast literature reporting single-cell recording data (for review see Bowers, 2009), there are only a handful of studies that have characterized the responses of single units in PDP and related artificial neural networks, and even fewer studies that have compared these results to the neuroscience findings. And as far as we are aware, this is the first study to compare the response properties of localist representations and single neurons. What is perhaps surprising is that when these comparisons are made the results highlight the biologically plausibility of localist coding schemes.

Two findings from neuroscience are often taken in support of selective distributed rather than grandmother cell coding schemes; namely, the fact that selective units often respond to more than one category of input, and the frequency with which selective neurons are found (Waydo et al., 2006). Although we agree that these findings are problematic for grandmother cell theories when grandmother cells are defined extremely narrowly (that is, the view that a single neuron fires in response to a given input and responds at baseline levels to everything else), our simulations with the SCM model of visual word identification show that localist representations can accommodate these two sets of results. Accordingly, the neuroscience results should not be used to reject localist models in favor of PDP models in psychology.

Furthermore, recent analyses of single units in PDP models also lend some support to localist coding schemes. That is, in contrast with the widespread view that PDP models learn highly distributed codes (a key reason why so few researchers have analysed single hidden units one-at-a-time), these models sometimes learn localist representations (Bowers et al., 2014, 2016). Even "deep" networks that can match human performance on some object recognition tasks often learn highly selective codes (e.g., Yosinski, Clune, Nguyen, Fuchs, & Lipson,

2015). In our view, the key message for psychologists is that analyses of single units in both PDP and localist models demonstrate the biological plausibility of localist representation. The message for neuroscientists is that theorists should consider the hypothesis that brain implements localist coding schemes in which single neurons represent one thing, but fire to multiple things. Indeed, Bowers (2009) suggested grandmother cells be defined as localist representations. Adoption of this definition would ensure that psychologists and neuroscientists were using a common language, and no one would be dismissing a hypothesis that few if any researchers take seriously (or indeed, ever proposed).

Review of Results

We now consider the present results in some more detail. In Simulation 1 we carried out single unit recordings from the word units in the original SCM without changing any of the parameters. Here we found that indeed the degree of selectivity was higher than estimated by Waydo et al. (2006). That is, rather than a selectivity estimate of .5% we found .01%. This might at first appear to support the common claim that the neuroscience is inconsistent with localist coding models.

There are two reasons why we think this conclusion is unjustified. First, Waydo et al. (2006) note themselves that their estimate of .5% selectivity may be a great overestimate given the technical and methodological limitations of the Quiroga et al. (2005) study. They write:

> Two significant factors may bias our estimate of sparseness upward. A large majority of neurons within the listening radius of an extracellular electrode are entirely silent during a recording session (e.g., there are as many as 120–140 neurons within the sampling region of a tetrode in the CA1 region of the hippocampus (Henze et al., 2000), but we typically only succeed in identifying 1–5 units per electrode)...Thus, the true sparseness could be considerably lower. Furthermore, there is a sampling bias in that we present stimuli familiar to the patient (e.g., celebrities, landmarks, and family members) that may evoke more responses than less familiar stimuli. For these reasons, these results should

> be interpreted as an upper bound on the true sparseness, and some neurons may provide an even sparser representation.

That is, the authors are suggesting that their estimate of sparseness (meaning selectivity, cf. Footnote 1) may be off by more than an order of magnitude. In which case, the level of selectivity we observed in original model is not outside the range of possible selectivity values.

Second, the parameter settings of the original SCM model were not constrained in any way by estimates of neural selectivity. This raises the question of whether a modified version of the model can better capture the neural data as reported. In fact, there are good behavioral data to suggest that the level of selectivity is too extreme in the published SCM model. In particular, there is good evidence that a large cohort of words is co-activated when a person identifies a single word. For instance, while processing the meaning of a given word (e.g., PEAR or WARM) the meaning of form related words (e.g., BEAR or ARM) are also activated (Bowers, Davis, & Hanley, 2005; Pecher, de Rooij, & Zeelenberg, 2009). Given that a specific target word shares overlap many different words (e.g., PEAR is closely related to BEAR, DEAR, FEAR, HEAR, NEAR, REAR, TEAR, WEAR, YEAR, EAR, PEA, etc. and it overlaps with many more words to a lesser degree), this strongly suggests that a large cohort of form related words are co-activated at the same time. This is not captured in the original SCM model.

In Simulation 2 we modified the way lateral inhibition works in the model, such that the activity of a word unit is used to weight the lateral inhibitory input to that unit. This reduces winner-take-all behaviour by allowing weakly activated units to increase their activity in response to positive input without being immediately suppressed. This type of

shunting inhibition is used in other localist models, notably in those studied by Grossberg and colleagues (e.g., Carpenter & Grossberg, 1980; Grossberg, 1973). The key result of this modification to the model is that it is possible to manipulate level of selectivity by varying a single parameter. By choosing an appropriate value for this parameter we are able to match the level of selectivity estimates reported by Waydo et al. (2006). Importantly, the model is still able to identify all the words in its vocabulary. Of course, in order to make any claims regarding the viability of the modified model as a theory of visual word recognition it will need to account for all the behavioral data that the original model can explain. But this is not the goal of the present paper. Indeed, we do not intend to link the viability of localist coding to this specific model of word identification. Rather, we simply use this demonstration as an existence proof that a localist model that can identify a large set inputs (in this case over 30,000 words) can straightforwardly capture the selectivity data that are often used to reject grandmother cells in neuroscience and localist representations in psychology.

One possible objection to our simulations above is that we were randomly sampling from ~30,000 word units (in which every unit represents a word) whereas in neuroscience studies researchers are randomly sampling from many millions of neurons (in which many neurons may not even be involved in the identify a stimulus). The different conditions may lead to very different estimates of selectivity. For example, our selectivity measures observed in Simulation 2 would have been much greater (and inconsistent with the observed data) if we embedded the 30,000 units amongst millions of other units that did not fire to the inputs).

The problem with this critique is that it rests on the assumption that there is a single neuron/unit that codes for a specific word. But this is not the assumption of any biologically plausible theory in neuroscience or psychology. Indeed, Barlow (1985), Gross (2002), Page (2000), and Perrett et al. (1989), among others, are all clear that redundant coding would be

required in any feasible grandmother coding scheme, and level of redundancy might scale with the familiarity of the item (Konorsky, 1967). Similarly, we imagine that there might be many 1000s of redundant units in any biologically plausible localist model. One (simplistic) way to simulate this for present purposes would be to simply replicate the SCM model 1000 times (so that there are 1000 redundant representations for each word), such that there are 30,000 x 1000 units, or 30 million word units. This scaled-up model would result in identical selectivity values. Perhaps another 30 million units could be added that are not currently committed to any words, making a network of 60 million units with a selectivity value of 50% of our estimate. And many more units might be considered that code for sublexical representations (e.g., letters, bigrams, etc.) that fire to more inputs, which in turn might bring the selectivity values back closer to our original estimates. The point is simply that it is possible to scale up the current model to include many millions of units while maintaining similar levels of selectivity.

To continue with this (admittedly speculative) analysis comparing number of units in this highly redundant SCM network and brain areas, how many neurons might there be in the visual word form area (the brain area one would record from if you were looking for selective responses to words)? It has been estimated that there are ~100,000 cells per cubic mm in cerebral cortex (Braitenberg & Schüz, 1980), and the volume of the visual word form area may be roughly ~175 cubic mms (based on the number of active voxels; Baker, Liu, Wald, Kwong, Benner, & Kanwisher, 2007). This would suggest that there are approximately 17.5 million cells in the visual word form area. Of could this estimate could be way off, but it does at least illustrate that a redundant version of the SCM model that includes 1000 units associated with each word may have as many (or more) units than neurons in the visual word form area. There have not been single-cell recording studies carried out in the visual word

form area in humans, but if localist model of word identification is to be maintained, we would have to predict that similar levels of selectivity are obtained when recording from cortical areas devoted to coding words.

Quian Quiroga et al. (2005) carried out their single-cell recording studies in the hippocampus (and related structures) which includes many times more neurons than visual word form area (perhaps an order of magnitude or much more). But there are also good reasons to think that episodic memories are coded in a much more redundant manner. For instance, according to the multiple trace hypothesis (Moscovitch & Nadel, 1998), a separate memory trace is encoded each time a given items is encountered and remembered (e.g., each time you see a new episode of the TV show *Friends* new traces of the actress Jennifer Aniston are stored) and new memory traces are encoded each time you remember of something (e.g., every time you think about Jennifer Aniston will store new memory traces of Jennifer Aniston). With massive redundancy, similar levels of selectivity might be expected even when randomly probing for localist representations amongst 100s of millions of neurons. In addition, putting aside any comparisons in the number of neurons in various cortical regions and models, Waydo et al. (2006) estimated that each neuron they recorded from responded to between 50-150 different images (based on the assumption that people recognize between 10,000 and 30,000 discrete objects). We found that our localist units responded to a similar number of words (based on a model that knows ~30,000 words). Accordingly, we would argue that our method of analyzing the SCM model provides a plausible comparison to single-cell recording studies reported by Waydo et al. (2006), and as a consequence, results from such studies should not be used to rule out localist coding models.

REFERENCES

Adelman, J. S., Gubian, M., & Davis, C. J. (2016). *The easynet software for simulating models of visual word recognition*. Manuscript in preparation.

Baker, C. I., Liu, J., Wald, L. L., Kwong, K. K., Benner, T., & Kanwisher, N. (2007). Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proceedings of the National Academy of Sciences, 104(21*), 9087-9092. doi:10.1073/pnas.0703300104

Barlow, H. B. (1985). The 12th Bartlett Memorial Lecture: The role of single neurons in the psychology of perception. *Quarterly Journal of Experimental Psychology: Section A. Human Experimental Psychology, 37*, 121–145.

Bell, D., Forster, K., & Drake, S. (2015). Early semantic activation in a semantic categorization task with masked primes: Cascaded or not? *Journal of Memory and Language, 85*, 1-14. doi:10.1016/j.jml.2015.06.007

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"?. *Journal of Memory and Language*, 52(1), 131-143. doi:10.1016/j.jml.2004.09.003

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological Review, 121*, 248-261. doi:10.1037/a0035943

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2016). Why do some neurons in cortex respond to information in a selective manner? Insights from artificial neural networks. *Cognition, 148*, 47-63. doi:10.1016/j.cognition.2015.12.009

Braitenberg, V., and A. Schüz. 1998. Cortex: Statistics and Geometry of *Neuronal Connectivity. 2d ed.* Berlin: Springer. doi:10.1007/978-3-662-03733-1

Carpenter, G.A. & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37(1), 54-115. doi:10.1016/S0734-189X(87)80014-2

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117*, 713-758. doi:10.1037/a0019738

Eichenbaum, H. (2001). Engram. In P. Winn (Ed.), *Dictionary of biological psychology* (p. 558). New York: Routledge

Flusberg, S. J. & McClelland, J. L. (2014). Connectionism and the emergence of mind. S. Chipman (Ed.), *The Oxford Handbook of Cognitive Science.* Forthcoming: published on-line, Nov 2014

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics, 52,* 217–257. doi:10.1002/sapm1973523213

Gur, M. (2015). Space reconstruction by primary visual cortex activity: a parallel, non-computational mechanism of object representation. *Trends in Neurosciences, 38*(4), 207-216. doi:10.1016/j.tins.2015.02.005

Konorski, J. (1967). Integrative activity of the brain; an interdisciplinary approach. Chicago: University of Chicago Press.

Lupker, S. J., Zhang, Y., Perry, J. R. & Davis, C. J. (2015). Superset versus substitution-letter priming: An evaluation of open-bigram models. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 138-151. doi:10.1037/a0038392

Lytton, W. W. (2007). *From computer to brain: foundations of computational neuroscience*. Springer Science & Business Media.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. Psychological Review, 101, 3-33. doi:10.1037/0033-295X.101.1.3

McClelland, J.L. & Ralph, M.A.L. (2015). Cognitive Neuroscience. In: J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 4* (pp. 95-102). Oxford: Elsevier, Chicago. doi:10.1016/b978-0-08-097086-8.56007-3

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: 1. An account of basic findings. *Psychological Review, 88,* 375– 407. doi:10.1037/0033-295X.88.5.375

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing: Psychological and biological models (Vol. 2).* Cambridge, MA: MIT Press.

Moscovitch, M., & Nadel, L. (1998). Consolidation and the hippocampal complex revisited: in defense of the multiple-trace model. *Current Opinion in Neurobiology, 8(2),* 297-300. doi:10.1016/S0959-4388(98)80155-4

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences, 2*, 455– 462. doi:10.1016/S1364-6613(98)01241-8

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23*, 443–512. doi:10.1017/S0140525X00003356

Pecher, D., de Rooij, J., & Zeelenberg, R. (2009). Does a pear growl? Interference from semantic properties of orthographic neighbors. *Memory & cognition*, 37(5), 541-546. doi:10.3758/MC.37.5.541

Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., et al. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology*, 146, 87–113.

Polk, T. A., & Farah, M. J. (2002). Functional MRI evidence for an abstract, not perceptual, word-form area. J*ournal of Experimental Psychology: General, 131(1*), 65-72. doi:10.1037/0096-3445.131.1.65

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005, June 23). Invariant visual representation by single neurons in the human brain. *Nature, 435*, 1102–1107. doi:10.1038/nature03687

Rolls, E. T., Treves, A., & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Experimental Brain Research*, 114, 149 –162. doi:10.1007/PL00005615

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations. MIT Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523-568. doi:10.1037/0033-295X.96.4.523

Stinchcombe, E. J., Lupker, S. J., & Davis, C. J. (2012). Transposed-letter priming effects with masked subset primes: A re-examination of the "Relative position priming constraint". *Language and Cognitive Processes*, *27*, 475-499. doi:10.1080/01690965.2010.550928

Vankov, I.I. & Bowers, J.S. (this issue). Learning localist representations in feed-forward Parallel Distributed Processing networks.

Waydo, S., Kraskov, A., Quiroga, R. Q., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. J*ournal of Neuroscience*, *26*, 10232–10234. doi:10.1523/JNEUROSCI.2101-06.2006

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, *16*(8), 487-497. doi:10.1038/nrn3962

Footnotes

Footnote 1: Somewhat confusingly, Waydo et al. 2006 used the term sparseness to refer to neural selectivity; that is, they defined sparseness as the proportion of stimuli a neuron responds to. Here we use the term selectivity, and reserve the term sparseness to refer to the proportion of neurons in a population of neurons that fire in response to a given image. Although these two measures are closely related, the two measures can dissociate (see Bowers, 2011; Foldiak, 2009)

Footnote 2: http://adelmanlab.org/easyNet/

Footnote 3: All simulations reported in this article can be reproduced in *easyNet* by downloading the material (input data and code) available at: http://adelmanlab.org/easyNet/downloads/Shunted-SCM_files/

Footnote 4: As described in Davis (2010), the model computes match values (varying between zero and one) which reflect the similarity of the input stimulus to each of the words contained in the model's vocabulary. The α parameter scales the strength of this excitatory input to the word layer.